8 Genome Sequencing of *Capsicum* Species: Strategies, Assembly, and Annotation of Genes

Pasquale Tripodi, Alberto Acquadro, Sergio Lanteri and Nunzio D'Agostino

Abstract

Pepper (*Capsicum* spp.) belongs to the *Solanaceae*, which is an economically important family of flowering plants consisting of about 102 genera and over 2500 species. The *Solanaceae* family includes crops of agronomic importance for which the efforts in genome sequencing are ongoing by almost 10 years (https://www.solgenomics.net/organism/sol100/view). Since the beginning of 2014, various consortia have released the genome sequences of domesticated and wild *Capsicum* species. The first effort was focused on the whole-genome sequencing of *Capsicum annuum* CM334 and of *Capsicum chinense* PI159236, which were widely used as founders of mapping populations and carry important disease resistance traits. Just a couple of months later, the genome sequences of *C. annuum* Zunla-1 and of the wild species Chiltepin (C. *annuum* var. *glabriusculum*) were published. Both studies reported a pepper genome size of ~3–3.5 Gb, rich in repetitive elements (over 80%) with about 35 thousand genes. The improved version of the reference genome CM334 as well as of *C. chinense* PI159236 together with the sequencing of the domesticated *Capsicum baccatum* revealed evolutionary relationships and estimated lineage divergence times occurring in *Capsicum*. Recently, the linked-read sequencing technology has been applied for the sequencing of a *C. annuum* accession that was an F_1 cross hybrid of CM334 and a non-pungent pepper breeding line. Furthermore, genome resequencing studies have been performed with the aim to analyze *loci* of interest related to biotic/abiotic stresses and to qualitative features. In this chapter, we provide an overview of the genome sequencing and annotation strategies and describe the main results disclosed by all the whole and targeted genome sequencing projects in *Capsicum*.

8.1 International Initiatives in Pepper Genome Sequencing

Since the release of the first whole-genome sequence of a plant species (Arabidopsis Genome Initiative 2000), various national and international initiatives have led to the sequencing and assembling of genomes of both crop and non-crop plants from different clades. During the last five years, transnational research consortia released the genome sequences of domesticated and wild peppers. The main information arising from *Capsicum* spp. genome sequencing has been obtained by two international groups: the

P. Tripodi (✉) · N. D'Agostino
CREA Research Centre for Vegetable and Ornamental Crops, Pontecagnano Faiano, Italy
e-mail: pasquale.tripodi@crea.gov.it

A. Acquadro · S. Lanteri
DISAFA, Plant Genetics and Breeding, University of Torino, Turin, Italy

N. Ramchiary and C. Kole (eds.), *The Capsicum Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-319-97217-6_8

former composed by scientists of 27 institutes from Korea, Israel, and USA (Kim et al. 2014), the latter by researchers of 13 institutes from China, Mexico, France, and USA (Qin et al. 2014). Since then, two other complete genome sequences have been released into the public domain (Kim et al. 2017; Hulse-Kemp et al. 2018). Descriptions of the genome sequencing and annotation strategies as well as the main results disclosed are reported below.

8.1.1 The Genome Sequence of Hot Pepper

In January 2014, Kim et al. (2014) reported the whole-genome sequencing and assembly of the Mexican landrace *C. annuum* cv. Criollo de Morelos 334 (hereafter CM334) and the *Capsicum chinense* accession PI159236 as the foundation for interspecies comparative analysis (Table 8.1). Both accessions were selected for being resistant to diseases including *Phytophthora* spp., Nematodes, *Tobacco Mosaic Virus* (TMV), *Potato Virus Y* (PVY), *Tomato Spotted Wilt Virus* (TSWV), *Pepper Mottle Virus* (PepMoV), *Tobacco Etch Potyvirus* (TEV). The authors also accomplished the resequencing of two cultivated peppers: *C. annuum* cv. 'Perennial' and *C. annuum* cv. 'Dempsey', which were the parents of a 120 recombinant inbred F_8 line (RIL) population used for the development of high-density genetic and physical maps. Paired-end (PE) and mate-pair (MP) libraries were sequenced on Illumina platforms (GAII and HiSeq 2000). As for CM334, a total of 650.2 Gb of genome sequence (coverage 186.6×) was generated from genomic libraries with insert sizes ranging from 180 bp to 20 Kb and read lengths of 36, 76, or 101 bp. In case of *C. chinense*, a total of 289.6 Gb of genomic sequence data was generated. Prior to the genome assembly, an *in-house* preprocessing pipeline was adopted to filter out low-quality sequences for short reads, by eliminating contamination from publicly available bacterial genome sequences (identity > 98%, coverage > 50%), duplicated reads, low-quality sequences as well as correcting substitution sequencing errors. A 19-mer analysis was performed to determine the genome size of CM334 and PI159236, which were estimated to be 3.48 and 3.14 Gb, respectively. To generate initial contigs, all the reads from both CM334 and *C. chinense* libraries were first merged to single reads ignoring pair information and then assembled into 37,989 scaffolds with N_{50} (the average length value of fragments for the 50% of the genome) of 2.47 Mb. The assembled CM334 genome sequence was validated with 27 bacterial artificial chromosomes (BACs) from euchromatic/heterochromatic regions with insert size larger than 70 Kb. All scaffolds matched the complete BAC sequences with more than 99.9% identity and 26 BACs were covered by single scaffolds. The quality of PI159236 genome assembly was assessed on the basis of about 600 *C. chinense* expressed sequence tags (ESTs) and mRNAs and additional 35,000 annotated CM334 genes, of which 97% matched with the PI159236 assembly. The validation confirmed about 23,000 genes in the *C. chinense* genome and made it possible the identification of a core gene set shared by the two accessions (Kim et al. 2014). To support the scaffolding process and construct pseudomolecules, a high-density genetic map was generated through low-depth (1×) whole-genome sequencing of 120 intraspecific *C. annuum* RILs ('Perennial' × 'Dempsey'). Over 3 million single nucleotide polymorphism (SNP) markers were identified among CM334 and the parents of the RILs, and a set of 21,121 markers were selected for map construction (Kim et al. 2014). The final map consisted of 12 linkage groups and 6281 markers covering 3796 cM. A subset of 4562 markers (73%) allowed to anchor 86% of scaffolds (2.63 Gb; 1357 scaffolds) to 12 pseudo-molecules. Accuracy of the linkage map was validated using the conserved ortholog set II (COSII) maps previously developed (Wu et al. 2009). Resequencing data revealed the presence of 10.9 and 11.9 million divergent SNPs in 'Perennial' and 'Dempsey', respectively, when compared with the CM334 reference sequence. Sequencing of *C. chinense* highlighted 56.6 million SNPs compared to CM334. As a result, 94.5, 94.3, and 89.6% of the CM334 genome was covered by 'Perennial', 'Dempsey' and *C. chinense* sequences, respectively. Transposable elements (TEs), which have played a key role in shaping the DNA landscape of genomes during evolution and led to the conversion of euchromatin into heterochromatin, were found to represent a preponderant portion on the whole-genome in respect to tomato as well as other sequenced *Solanaceae* genomes, being the 76.4% (i.e., 2.34 Gb) in CM334 and 79.6% (i.e., 2.35 Gb) in *C. chinense*. TEs were widely dispersed throughout the pepper genome and their distribution was inversely correlated with gene density. The most frequent TEs were long terminal repeat (LTR) elements, representing more than 70% of the identified TEs in the two genomes. The composition of these repetitive DNA sequence motifs widely differs from the one detected in other crops, as the *Gypsy* elements were 12-fold more than the *Copia* elements and their proliferation caused the expansion of the pepper genome. On the other hand, the accumulation of *Copia* and *Tat* elements (a subgroup of the *Gypsy* clade) was responsible for the expansion of hot pepper euchromatin. A consensus annotation of 34,903 protein-coding genes was generated for CM334 (Pepper Genome Annotation v. 1.5). Over 93% of the predicted protein-coding sequences was supported by ~20 Gb Illumina RNA-seq data from four tissues/organs (flower, root, leaf, and fruit) at different stages of plant development. Furthermore, 177 microRNAs, corresponding to 37 microRNA families, were identified in CM334. The number of annotated genes was similar to the one previously identified in tomato and potato. Overall, the authors reported 23,245 hot pepper genes distributed in 16,345 families. By comparing pepper and tomato genomes, it was possible to identify 17,397 orthologous genes, whose expression was investigated through RNA-seq from tissues at the same developmental stage of the plants. In both crop species, it was possible to identify a high number of differentially expressed genes (DEGs) in pericarp and placenta (34 and 24.7% on average, respectively) while in root and leaf the number of DEGs was relatively low (15.1 and 8.8%, respectively). The distribution of orthologous gene families of six crops (hot pepper, tomato, potato, Arabidopsis, grape, and rice) allowed to identify 7826 shared gene families and 756 unique families to hot pepper. Furthermore, variations in family size were found in many hot pepper gene families, such as those involved in disease resistance and

Table 8.1 Comparison of the main features of the *Capsicum* sequenced genomes

	C. annuum CM334 (v1.55)[a]	*C. chinense* PI159236[a]	Zunla-1[b]	Chiltepin[b]	*C. baccatum* PBC81[c]	*C. chinense* PI159236[c]	CM334 F$_1$[d]
Total sequence length (Gb)	650.2	289.6	477.37	295.85	526.7	425.7	104.7
Sequencing depth (X)	186.6	83.2	146.43	96.37	136.1	132.2	56
Genome size (Gb)	3.48	3.14	3.26	3.07	3.9	3.2	3.2
Scaffold no	37.989	239.495	28,149*	30,293*	2.083	1.557	83.391
TE elements %	76.4	79.6	80.9	81.4	85	85	*na*
Genes number	34.903	33.788	35.336	34.476	35.874	35.009	*na*

[a] Kim et al. (2014)
[b] Qin et al. (2014)
[c] Kim et al. (2017)
[d] Hulse-Kemp et al. (2018)
* >2 K bp
na not available (annotation not provided)

cellular functions (i.e., *cytochrome P450* and heat shock protein genes).

8.1.2 The Genome Sequence of Cultivated and Wild Peppers

Few months after the release of the CM334 genome, Qin et al. (2014) published the reference genome sequences of the cultivated pepper Zunla-1 (*C. annuum* L.) and its wild progenitor Chiltepin (*C. annuum* var. *glabriusculum,* also termed *C. annuum* var. *aviculare*) (Table 8.1). Zunla-1 is an F_9 inbred line derived from a cross between two *C. annuum* cultivars grown by small farmers in China, while Chiltepin is a landrace collected in the north-central Mexico.

Eleven (6 PE and 5 MP) and nine (5 PE and 4 MP) Illumina libraries with different insert sizes were prepared for Zunla-1 and Chiltepin, respectively. These libraries were sequenced using the Illumina Genome Analyzer II device and generated 477.37 Gb (146.43× coverage) of raw sequencing data for Zunla-1 and 295.85 Gb (96.37× coverage) for Chiltepin.

After filtering out low-quality and duplicate reads, a total of 325.29 Gb (99.78× coverage) of high-quality sequence data for Zunla-1 and of 204.86 Gb (66.73× coverage) for Chiltepin was retained. For Zunla-1, after filling the gaps, the total scaffolds size was ~3.35 Gb (N_{50} = 1.22 Mb) and the total contig size was ~3.21 Gb (N_{50} = 55.43 Kb). For Chiltepin, after gap filling, the total scaffold size was ~3.48 Gb (N_{50} = 444.59 Kb), while the total contig size was ~3.3 Gb (N_{50} = 52.23 Kb). Based on an intraspecific *C. annuum* F_2 population, a high-resolution genetic map with 7657 SNP markers was generated and used to anchor and orient 4956 scaffolds from Zunla-1 to the 12 chromosome pseudo-molecules. Overall, 78.95% of the assembly (~2.64 Gb; 1822 scaffolds) was successfully anchored to the 12 pseudo-chromosomes. The unplaced 3134 scaffolds (705 Mb in total) were assigned to a pseudo-chromosome designated as ‘00’. By comparing the genetic and the physical distances, similar patterns of recombination were detected, which were markedly reduced in broad pericentromeric regions and consistent at chromosome ends. LASTZ (Large-Scale Genome Alignment Tool) (Harris 2007) was used to align the assembly of Chiltepin chromosomes to the Zunla-1 reference genome. The completeness and quality of the assemblies were evaluated by aligning pepper ESTs available at dbEST (https://www.ncbi.nlm.nih.gov/dbEST) as well as Illumina reads generated from short insert size libraries onto Zunla-1 and Chiltepin genomes, respectively. Similarly to what previously reported for the CM334 genome (Kim et al. 2014), more than 81% (~2.7 Gb) of the Zunla-1 and Chiltepin genomes is composed by transposable elements (TEs), most of which are LTR retrotransposons of the *Gypsy* clade (54.5%) followed by *Copia* (8.6%). Divergence analysis allowed to date the insertion time of LTRs ~0.3 million years ago (Mya), suggesting that the expansion of the pepper genome was quite recent during the evolution of the *Solanaceae* family.

In total, 35,336 and 34,476 protein-coding genes were predicted with high confidence in Zunla-1 and Chiltepin, respectively. Furthermore, over 90% of predicted genes were supported by different items of evidence (ESTs; RNA-seq data; homologous proteins). Gene discovery and annotation benefited from the generation of 30 RNA-seq libraries (over 90 Gb of sequence data) from various tissues/organs at different developmental stages. RNA-seq expression profiles highlighted constitutively expressed (over 31%) as well as tissue-specific genes. Discovery and annotation of long noncoding RNAs (lncRNAs) as well as of short interference RNAs (siRNAs) and microRNAs (miRNAs) in Zunla-1 was also performed by the RNA-sequencing of a flower bud library and five small RNA libraries from different tissues. Over 6500 lncRNAs, 5500 phased siRNAs and 176 miRNAs were identified. A set of 141 miRNAs were in common with other *Solanaceae*, while 35 miRNAs were classified as pepper-specific. Over 1100 target genes were identified, mostly coding for transcription factors, of which 78% have putative functions.

8.1.3 The Genome Sequence of *C. baccatum* and *C. chinense*

Recently, researchers of the consortium which previously released the CM334 genome, performed the sequencing and assembly of the genome of *C. baccatum* PBC81 (hereafter, *Baccatum*) and provided an improved version of the reference genome of both CM334 and *C. chinense* PI159236 (hereafter, *Chinense*) (Kim et al. 2017; Table 8.1).

The Illumina HiSeq 2500 platform was used for the sequencing of libraries with insert sizes in the range of 200 bp–10 Kb. In total, 526.7 Gb (136.1× coverage) and 425.7 Gb (132.2× coverage) of the *Baccatum* and *Chinense* genomes were generated. On the basis of 19-mer analysis, the estimated genome sizes were 3.9 and 3.2 Gb, respectively. For scaffold anchoring, high-genetic density maps were developed following genotype by sequencing of an F_2 *C. baccatum* intraspecific population (obtained by crossing lines 'Golden-aji' and 'PI594137') as well as segregating interspecific populations obtained by crossing *C. annuum* and *C. chinense*. The assembled genomes of *Baccatum* and *Chinense* were organized into 12 chromosomes-scale pseudo-molecules, being 3.2 and 3.0 Gb in size with scaffold N_{50} of 2.0 and 3.3 Mb, respectively. The total length of successfully anchored scaffolds were 2.8 Gb in (2083 scaffolds) for *Baccatum* and 2.8 Gb (1557 scaffolds) for *Chinense*, accounting for the 87 and 89% of the pepper genome, respectively.

As expected, repeated sequences represented the 85% of the entire genome and, in each species, over half was made up of LTR retrotransposons of the *Gypsy* clade. In the *Baccatum* genome, *Athila* elements were found to be more abundant (>two fold) and contributed to species-specific genome expansion in the *C. baccatum* lineage. On average, about 35,000 genes were annotated in both the *Baccatum* and *Chinense* genomes. In addition, a comparison between the updated and previous protein-coding gene annotation of CM334 revealed differences in ~10,000 gene models, most of which were associated with TEs in the previous genome annotation.

The phylogenetic analysis on *Baccatum*, *Chinense* and CM334 revealed a first lineage divergence between *Baccatum* and a progenitor of the other two peppers at about 1.7 Mya, followed by divergence between CM334 and *Chinense* at 1.1 Mya. It is noteworthy that comparison between *Baccatum*, *Chinense* and CM334 disclosed important dynamic genome rearrangements involving translocations among chromosomes 3, 5, and 9 differentiating *C. baccatum* from the other two species.

8.1.4 Linked-Read Sequencing of Reference Genome

In 2018, the pioneering linked-read sequencing technology has been applied in *C. annuum* (Hulse-Kemp et al. 2018) and generated a highly ordered and more contiguous sequence assembly in respect to the available *C. annuum* reference genomes (Table 8.1). This technology was used to sequence a F_1 heterozygous individual from a cross between CM334 and a non-pungent blocky accession. The authors used Illumina HiSeq × Ten sequencer (10× Chromium technology) to produce 2 × 150 paired-end sequences (56× coverage). The Supernova Assembler (Weisenfeld et al. 2017) was used to resolve complex repeats and separate chromosomes based on haplotype information. It produced locally phased haplotype blocks, or pseudohaps, as output. In particular, two individual haplotypes were generated. With the aim of generating a reference assembly (hereafter UCD10X), a single pseudohap was utilized. Indeed, the pseudohap1 assembly was made up of 83,391 scaffold sequences for a total size of 3.21 Gb. Over 83% of the assembled sequence (~2.67 Gb) was anchored to the 12 chromosomes along with 541 Mb of unplaced sequence. The N_{50} was 123 Kb, 3.69 Mb and 227.2 Mb for contigs, scaffolds, and pseudo-molecules, respectively. The quality of the assembly was assessed by comparing the order of contigs with four high-density pepper genetic maps (three transcriptome-based and one genomic-based) and highlighted a concordant marker order. Furthermore, physical location of markers was also compared with the CM334 genome V1.55 pointing out that marker positioning in pericentromeric regions is more reliable in case of UCD10X. This assembly was, in the end, compared with the other publicly available pepper genomes (i.e. CM334V. 1.55, Zunla-1V. 2.0, Chiltepin V 2.0) in terms of length of scaffold sequences and overall size of the assembly. All the genome assemblies were comparable even if the quality within pseudo-chromosomes was variable, especially in heterochromatic regions. On the whole, although some regions were not accurately assembled with the linked-read library technology, the latter demonstrated to provide a valuable tool also for the de novo assembly of complex, highly repetitive, and heterozygous plant genomes.

8.1.5 Insight into Genome Expansion

Similar to what was observed in tomato (Tomato Genome Consortium 2012) and petunia (Bombarely et al. 2016), pepper is a paleohexaploid as its genome is the results of ancient triplication event. Since its speciation within the *Solanaceae* family, the pepper genome experienced no additional whole-genome duplication; however, its size is approximately four times than the one of tomato and threefold larger than the one of potato. Tomato and pepper genomes share syntenic blocks highly conserved and a high representation of LTR retrotransposons. The genome released by Kim et al. (2014) evidenced that pepper chromosomes highly expanded in both euchromatic and heterochromatic regions with respect to other *Solanaceae*. Most regions of the pepper genome are very rich in constitutive heterochromatin, which consists mostly of repetitive sequences and transposable elements. Comparison with the tomato genome suggested that the gene-rich regions near heterochromatin in tomato became heterochromatic regions in the pepper by accumulating repetitive sequences. In both species, the distribution of LTR retrotransposons was investigated highlighting a major representation of Del (Gypsy superfamily) and the existence of specific elements in pepper (Pseudovirus, Sire, CMR) and tomato (CoDi-D). Moreover, all repetitive elements in tomato were found in pericentromeric heterochromatin regions while in pepper, their distribution was observed in both heterochromatic and euchromatic regions.

Both Qin et al. (2014) and Kim et al. (2014) evidenced how the genes responsible of pungency synthesis underwent to duplication events, highlighting the existence of independent duplications in 13 gene families compared with *Arabidopis*, tomato and potato. Recently, retroduplication events (RTE) were described in NLR genes which are the major contributors of resistances in plants (Kim et al. 2017). The authors confirmed how RTE are common phenomena involved in the evolution of plants.

8.1.6 Gene Families

In pepper, 16,956 gene families were reported accounting for 22,885 genes identified among the predicted 34,447 protein-coding sequences (Kim et al. 2014). A similar number was found in the genome released by Qin et al. (2014), in which 16,770 families, including 26,444 genes out of 35,336 protein-coding sequences, were detected. Over two thousands transcription factors and transcriptional regulators (6.25% of predicted genes), which cluster in 80 families, were identified. The number was comparable with that of other plant species, although ABI3VP1 and RWP-RK families were most represented. Overall, 85% (1829) of TF genes were anchored to the pseudo-chromosomes with higher and lower concentration on chromosome 3 (257) and 10 (102), respectively. Among transcription factors, 73 families were represented by WRKY genes and 106 by NAC, both involved in plant development and defense mechanisms. Expression profiles of NAC genes evinced that about 30% of the genes were highly expressed during fruit developmental stages and 13% with the highest abundance in developing fruits.

One hundred and twenty-three genes belonging to the AP2/ERF superfamily (cellular responses, growth, and development) were identified. This family was mainly represented by different subfamilies including ERF (80%), AP2 (17%), RAV (1.6%).

The cytochrome P450s family was represented by 447 genes distributed in 9 of the 11 groups identified in land plants all involved in different metabolic tasks.

Other families of genes involved in developmental mechanism include the flowering truss gene family which was represented by 16 members responsible for flowering regulation and shoot architecture and the cuticle biosynthesis genes, which play a key role in preserving plants from various abiotic and biotic stresses regulating water and gas exchanges.

Phospatase families included serine/threonine classified into two groups: PPP (Ser/Thr-specific phosphoprotein phosphatase) and PPM/PP2C (magnesium dependent protein phosphatise).

Members of other gene families involved in growth, defense and physiological activities and including RNA-binding proteins, auxin Response Factors (ARFs), receptor-like kinases (867 genes), nucleotide binding site (684) and glycoside hydrolase gene families have been also identified within the pepper genomes.

8.2 Generalized Workflow for Genome Assembly, Structural and Functional Annotation

In Fig. 8.1, it is reported a generalized flowchart of the pepper genome assembly pipeline. The genome annotation pipeline included two phases: 'structural annotation' and 'functional annotation'. The former refers to the identification of DNA elements (e.g., repetitive elements, protein-coding genes, etc.) embedded in the genome, while the latter allows attaching biological information to these elements. Even if genome annotation pipelines differ in details, they share a core set of features and best practices (Fig. 8.1).

Prior to gene prediction, a thorough annotation of repetitive sequences in newly sequenced genomes is of utmost importance (Maumus and Quesneville 2016). To this end, a combination of de novo and similarity-based approaches was used for the identification and classification of repetitive DNA sequence motifs in the genome (Maumus and Quesneville 2016) (Fig. 8.2a). Protein-coding genes were predicted using ab initio gene finder tools in combination with comparative methods. The former is based on the identification of regions with coding potential and on the detection of signals within the DNA known as typical of gene structures; the latter relies on the use of homologous sequences (ESTs, mRNAs, RNA-Seq tags, proteins) to deduce gene structure. Among all forms of evidence, RNA-Seq tags have the greatest potential to improve the accuracy of gene annotations (Yandell and Ence 2012).

In case of ab initio gene prediction, an array of different gene finders was independently run to predict coding genes (Fig. 8.2b). Since ab initio gene finders need to be trained on a set of known genes, the first step was the construction of a training dataset (D'Agostino et al. 2007). To accomplish this task, available full-length cDNA and assembled RNA sequences were splice-aligned versus genome sequences.

As for comparative methods, protein-to-genome alignments as well as EST/RNA-seq tag-to-genome alignments laid the foundations to identify evidence-based gene *loci* and define gene structure. Previously identified transcripts and full-length proteins from pepper as well as available sequences from model or phylogenetically related species were used.

In the final step, ab initio gene predictions and diverse similarity-based evidence types were combined into consensus gene structures. This was performed using a 'combiner' algorithm in conjunction with manual curation of miss-annotated genes (Lewis et al. 2002) (Fig. 8.2b). The final set of gene annotations were further filtered to remove weakly supported genes.

Biological description (i.e., gene functions) was assigned to protein-coding genes based on BLAST similarity searches against UniProt (The

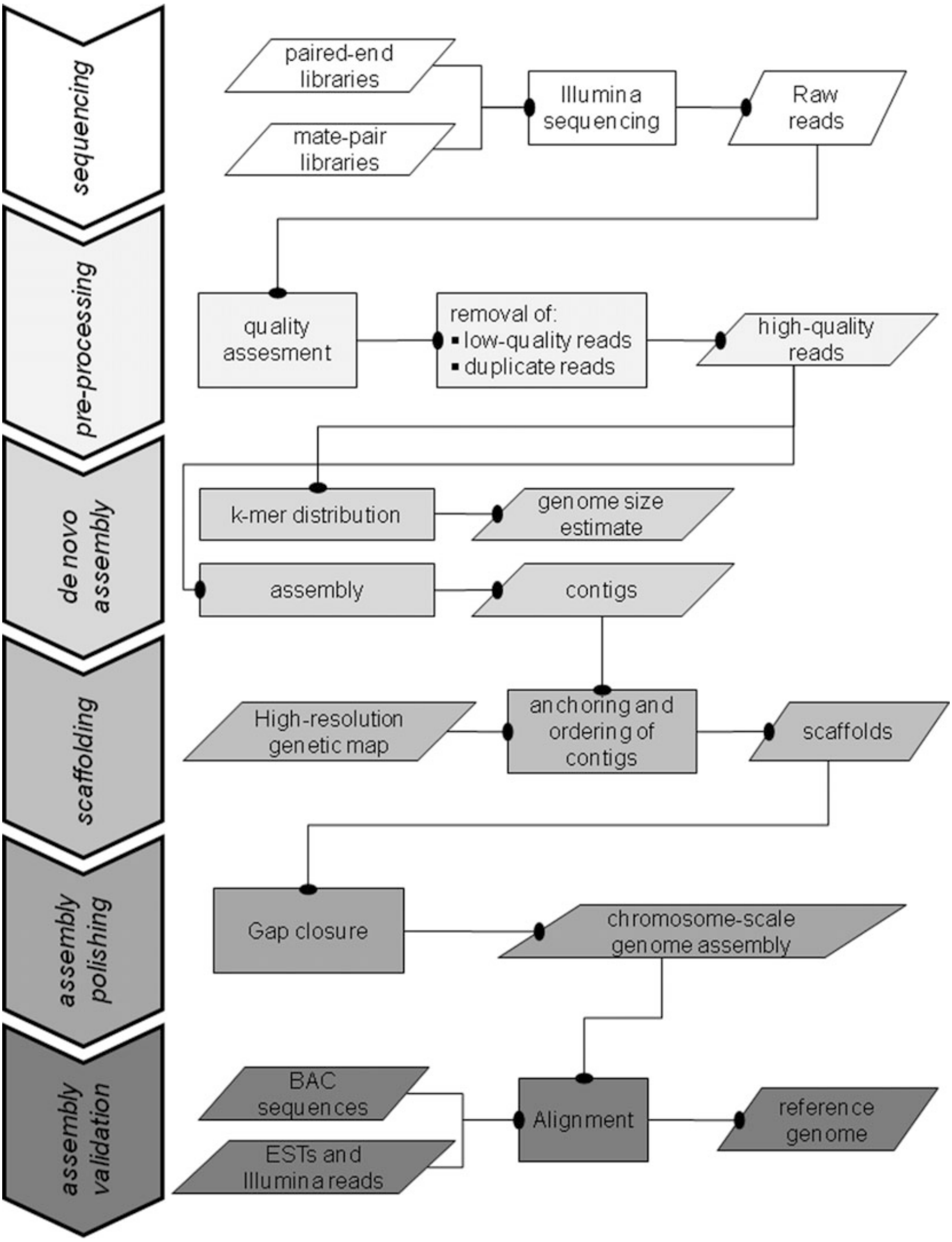

Fig. 8.1 Generalized flowchart of the pepper genome assembly pipeline. It can be divided into a core set of steps from the sequencing of Illumina paired-end and mate-pair libraries to the validation of the chromosome-scale genome assembly

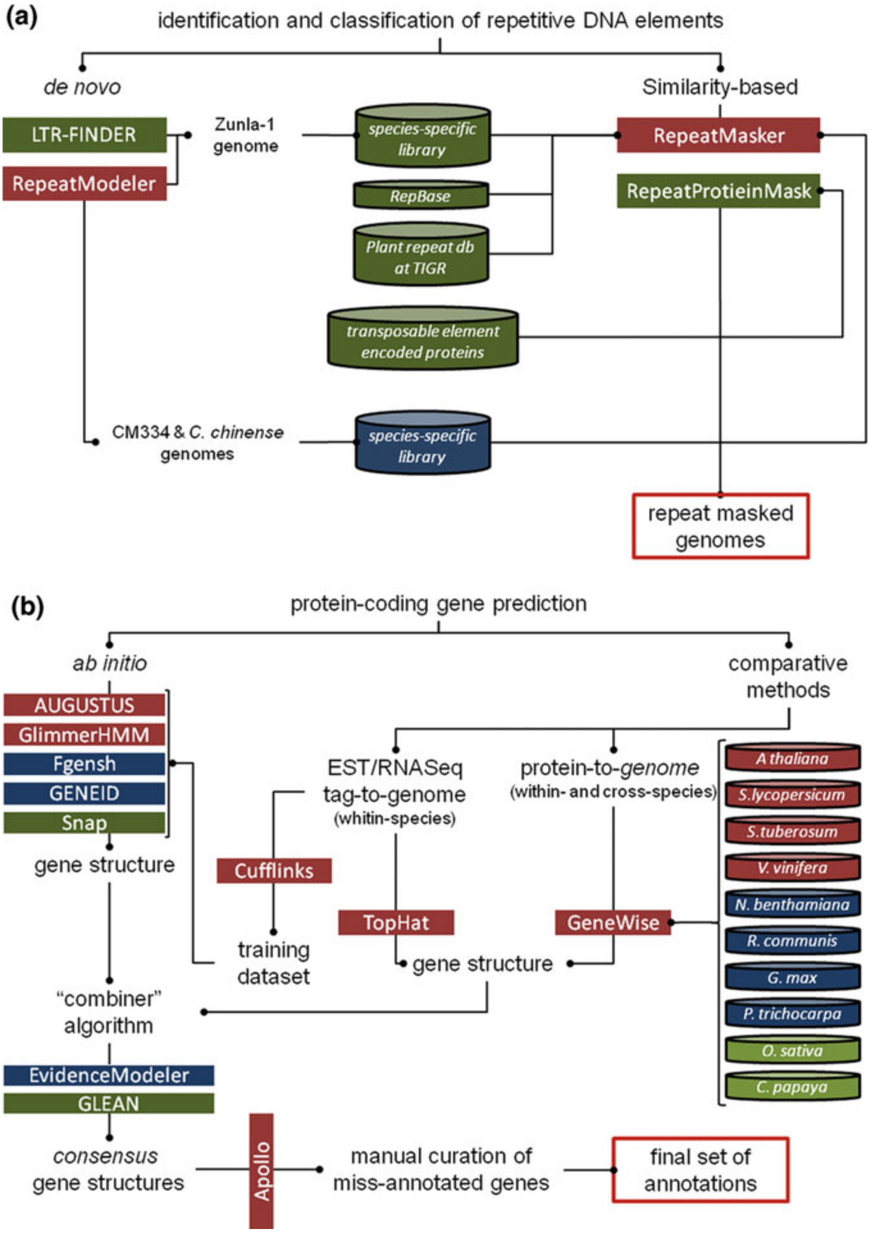

Fig. 8.2 Generalized flowchart of the structural genome annotation. Modules and corresponding software/tools are shown. In red are tools and databases common to both pipelines; in green those Zunla-1/Chiltepin specific; in blue those CM334 specific. **a** Identification and classification of repetitive DNA sequence motifs. **b** Prediction of protein-coding genes

UniProt Consortium 2015) and TAIR (Leonore et al. 2017) databases. In addition, motifs and domains within predicted protein sequences were identified using InterProScan (Jones et al. 2014). Sequence function description was integrated with gene ontologies (GO) (The Gene Ontology Consortium 2017) and the Enzyme Commission identifiers (EC; https://web.archive.org/web/20060218084611/; http://www.expasy.org/enzyme/) for a standard classification of gene products.

Where possible, the KEGG Orthology (KO) system was used to map proteins onto KEGG metabolic and signaling pathways (Kanehisa et al. 2017).

8.3 The Resequencing of *Capsicum* Genotypes

Thanks to the availability of pepper whole-genome sequences, NGS (next-generation sequencing) has been applied for the resequencing of additional pepper accessions in order to gather information on the structure, function, and evolution of genomes as well as to spot in detail allelic/structural variants.

To date, the resequencing of target genomic regions as well as of whole genomes of pepper accessions has been carried out with the goal to: (i) study how artificial selection traces embedded in the pepper genome correlates with breeding history, (ii) identify and fine mapping genomic *loci* conferring resistance against biotic stressors, (iii) reconstruct and structurally/functionally annotate the genomes of local pepper landraces.

8.3.1 Resequencing and Identification of Genes Involved in the Process of Pepper Domestication

Plant domestication is a complex evolutionary process in which human use of plant species led to morphological and physiological changes that distinguish domesticated taxa from their wild ancestors (Purugganan and Fuller 2009). The use of domestication as a model for the evolutionary process stems from an understanding of events associated with the origins of crop species and knowledge of the selective pressures experienced by domesticated taxa. The paper published by Qin et al. (2014) reported also the resequencing, at a coverage ranging from 20 to 30-fold depth, of 18 cultivated accessions representative the major varieties of *C. annuum* and two semi-wild/wild peppers, with the goal to provide insights into the identification of genes involved in the process of pepper domestication.

The alignment to the reference led to the identification of more than 9 M SNPs and 200 K small InDels, and both neighbor-joining tree and population structure highlighted that the wild and domesticated peppers are genetically distinguishable. To identify genomic footprints of artificial selection, a genetic bottleneck approach was used (Li et al. 2013). Genetic diversity was estimated by calculating θ_π (average pairwise divergence within a population) and θ_w Watterson's estimator, Watterson (1975). The regions showing significantly lower θ_π and θ_w in cultivars relative to the semi-wild/wild accessions were considered as potentially subjected to artificial selection. Only 2.6% of the genome, e.g., 115 regions containing 511 genes, appeared to be strongly affected by artificial selection in the cultivated peppers. The 511 spotted genes were mainly related to transcription regulation, stress and/or defense response, protein–DNA complex assembly, growth and fruit development, and ripening-associated biological processes. Among them 34 transcription factors (TFs), including activating protein (AP2), ethylene-responsive element-binding factor (ERF), and basic helix-loop-helix (bHLH) families as well as 10 disease resistance protein containing the NB-ARC domain were spotted. Those results show how resequencing approaches may contribute to the identification of genes related to morphological and physiological differences between cultivated and wild peppers and underlying pepper domestication and genetic improvement.

8.3.2 Resequencing Approaches for Genetic Analyses of Biotic Stress Resistance

The resequencing of *Capsicum* genomic regions and whole genomes has made it possible to identify molecular markers tightly linked to genes affecting resistance to biotic stresses and exploitable for marker-assisted breeding (MAS).

Devran and co-authors (2015) employed NGS technology in combination with bulk segregant analysis (BSA) for the identification of new molecular markers tightly linked to the Pvr4 *locus*, located on chromosome 10 and conferring dominant resistance to three pathotypes of Potyvirus (PVY, *Potato Virus Y*) as well as to *Pepper mottle virus* (PepMoV; Caranta et al. 1996, 1999). The susceptible *C. annuum* cv. 'SR-231' was crossed with the resistant accession 'Criollo de Morelos 334' (CM334) and a F_2 segregating progeny was generated from a single F_1 plant. The DNA of 15 resistant and 15 susceptible F_2 plants was at first pooled in two bulks, which were Illumina sequenced together with the parents. Due to the high synteny of tomato and pepper chromosome 10 (Wu et al. 2009), the sequence of tomato chromosome 10 was used as a reference for the alignment of Illumina reads of pepper parental lines, while reads from the two bulks were used to confirm the recognized polymorphisms. Some of the identified SNPs were then converted into CAPS (Cleaved Amplified Polymorphism sequence) marker, and the Pvr4 *locus* was mapped between the CAPS markers MY262 and MY69.

Thanks to the subsequent availability of the *Capsicum* genome sequences, pepper and tomato genomic regions including the Pvr4 *locus* were compared. A high degree of synteny was detected although the pepper chromosome 10 resulted inverted compared to tomato, and a tomato DNA region of approximately 1 Mb aligned against the corresponding pepper region that is three times wider. More of 5000 polymorphic sites (InDels and SNPs) were further spotted and markers developed. This allowed the fine mapping of *Pvr4* between two flanking markers (MY1176 and MY5009), with only one estimated recombination event on either side (less than 1 cM genetic distance away from the *locus*). The identification of two tightly linked flanking markers represents a highly reliable tool for easily transferring the Pvr4 *locus* to pepper breeding lines via marker-assisted backcrossing (MAB) selection.

Kang et al. (2016) performed the resequencing of the pepper cultivars 'YCM344' and 'Taean'. The former is highly resistant against Bacterial wilt (BW) which is caused by the soil-borne bacterium *Ralstonia solanacearum,* a pathogen distributed from tropical to temperate areas and which affects a broad range of dicot and monocot hosts, being particularly harmful for solanaceous crops. The two cultivars were Illumina resequenced at a coverage of 10× and the reads showed mapping rates higher than 93% to the CM334 reference genome (Kim et al. 2014). Approximately 7 K SNPs were detected in both accessions with frequencies ranging from 1.95 SNPs/Kb in 'Taean' to 2.01 SNPs/Kb in 'YCM334'.

The resequenced genomes were compared to each other with the goal to identify the most informative alleles related to BW resistance. More than 5, 6 M polymorphic SNPs and 149 K InDels were identified. This dataset allowed to identify genetic markers able to distinguish both these cultivars from CM334 and the two cultivars from each other. More than 100 K of the polymorphic SNPs were within gene regions, while ~36 K were in coding sequences (CDS), of which 23,396 showed non-synonymous (non-Syn) protein changes in 9102 genes.

Among the ten most polymorphic genes between 'YCM344' and 'Taean', two encodes for a 'Putative disease resistance protein' (CA10g15480 and CA12g20430) and were assigned to the 'Late blight resistance protein R1' gene family (IPR021929). This result suggested that the detected polymorphisms could be responsible for the different response to disease of the two cultivars. Other highly polymorphic genes included polyproteins, LRR like receptor kinases, N-like proteins, CC-NBS-LRR proteins, and putative phosphatidylinositol 4-kinase. A comparative analysis of SNPs located in genomic regions showing high similarity with known resistance genes in the tomato genomes was also performed. Among them, a total of seven genes showed non-Syn changes between 'YCM334' and 'Taean', which may be related to functional differences between the cultivars and represent strong candidate *loci* that contribute to BW disease in the cultivar 'YCM334'.

Recently Ahn et al. (2018), with the goal to discover SNPs associated with Powdery Mildew

(PM) resistance, resequenced via Illumina (∼11× coverage) the resistant *C. baccatum* line'PRH1' and the susceptible *C. annuum* line 'Saengryeg' (∼10× coverage). The agent of PM is *Leveillula taurica*, which is spread in a wide range of environments and represents a devastating fungal disease in pepper. The level of resistance to PM was assessed in both the lines, as well as in 45 individuals of their RIL F_4 population, through co-cultivation with powdery mildew and by using a scale ranging from one (resistant) to five (susceptible).

About 6 M SNPs, whose majority was classified as homozygous, were detected in both lines and found differentially distributed among the chromosomes. About 4.8 M SNPs were polymorphic between the two lines and were used for the design of 306,871 high-resolution melting (HRM) marker primer sets. The highest number of heterozygous SNPs was detected on chromosome 1 of PRH1 (i.e.: 23,932) and on chromosome 12 (i.e.,: 15,942) in 'Saengryeg', while the lowest one on chromosome 8 in both pepper lines (11,915 in 'PRH1' and 7229 'Saengryeg'). Based on their position within the genome sequence, the SNPs were then classified into intergenic or genic, and these latter sub-classified as intron SNPs, which were more frequent, and coding SNPs.

With the goal to gain deeper insight into SNPs associated with genes involved in disease resistance and stress tolerance processes, a chromosome wide functional annotation of the polymorphic variants among the two lines was performed. In introns and coding regions up to 6281 SNPs, associated with 46 RGA (Resistance Genes Analogues) carrying nucleotide binding site-leucine-rich repeat (NBS-LRR) motifs, were identified, found predominantly distributed on chromosome 4. NBS-LRR represents a large family of proteins that are encoded by RGA and are involved in pathogen recognition, including powdery mildew (Meyers et al. 2003; Coleman et al. 2009). Since the highest number of NB-LRR-linked SNPs was present in the PM resistant line 'PRH1' compared to the susceptible line 'Saengryeg', the authors assumed that NB-LRR resistance genes might play a key role in PM resistance. A subset of the identified SNPs was validated through HRM assay and, among the 36 primers applied, 19 significantly distinguished both parental lines and the resistant and susceptible plants in the F_4 progeny.

8.3.3 Resequencing of Pepper Landraces

Farmers' selection and adaptation to local climate and low-input agricultural practices has resulted in a plethora of pepper landraces that differ in growth habit, fruit shape and size, and organoleptic properties and that frequently carry resistance genes that are effective against abiotic and biotic stress. In the Piedmont region (northwest Italy) valuable and morphologically distinguishable landraces are grown, which are the result of a long selection process for adaptation to specific ecological niches. Thanks to the recent availability of *Capsicum spp.* genome sequences (Kim et al. 2014; Qin et al. 2014), Barchi and colleagues (2017) performed the genome resequencing of inbred lines of the four main landraces grown in the Piedmont region, namely: 'Cuneo' and 'Quadrato' (blocky types), 'Corno' (long type) and 'Tumaticot' (with small, sub-spherical fruits). The sequencing of the four genotypes was performed through Illumina technology, at coverage of ∼35X, and each genomic sequence was assembled into 12 chromosome-scale pseudo-molecules. Approximately 35 k genes were identified of which about 75% contained at least one IPR domain. The protein complements of the four reconstructed genomes, together with that of the reference (CM334), were analyzed to identify orthologs and orthogroups. More than 170 K sequences were clustered into 34,664 gene families (excluding singletons) of which 26,270 resulted to be shared among the five accessions, while only 152 gene families were in common between the two blocky types ('Quadrato' and 'Cuneo').

By aligning reads of the resequenced genotypes to the CM334 genome using standard pipelines, a set of about 19 M SNP/InDel was detected, ranging from 16.33 M ('Tumaticot') to

18.07 M ('Corno'). As expected for a selfing crop, the heterozygosity was rather low and ranged from ~0.2% in 'Corno' to ~0.1% in 'Tumaticot'.

A survey of the SNPs within genes that generally affect fruit size and shape in the *Solanaceae* (Chunthawodtiporn et al. 2018), were performed and mutations in the coding sequences of fw2.2, WUSCHEL (WUS) and fw3.2 were found to be common to the 4 genotypes while single deleterious mutation in sun-like ortholog gene was predicted. Differently, regulatory regions were rich in mutations with the exception of those related to the fw2.2 and fw3.2 *loci.* The large allelic diversity identified in the four resequenced accessions suggests that the pepper landraces under investigation can be considered as highly valuable pre-breeding resources.

References

Ahn YK, Manivannanbinaya A, Sandeep K, Jun TH, Yang EY, Choi S, Kim JH, Kim DS, Lee E-S (2018) Whole genome resequencing of *Capsicum baccatum* and *Capsicum annuum* to discover single nucleotide polymorphism related to powdery mildew resistance. Sci Rep 8:5188. https://doi.org/10.1038/s41598-018-23279-5

Barchi L, Acquadro A, Portis E, Comino C, Nourdine M, Borras D, Bustos Lopez M, Giordano R, Monge S, Carli C, Lanteri S (2017) Genome re-sequencing of piedmontese pepper ecotypes. In: The XIV *Solanaceae* and III cucurbitaceae genomics joint conference, 3–6 Sept, Valencia, Spain

Bombarely A, Moser M, Amrad A, Bao M, Bapaume L et al (2016) Insight into the evolution of the *Solanaceae* from the parental genomes of *Petunia hybrida*. Nat Plants 2(6):16074. https://doi.org/10.1038/nplants.2016.74

Caranta C, Palloix A, Gebre-Selassie G, Lefebvre V, Moury B, Daubeze AM (1996) A complementation of two genes originating from susceptible *Capsicum annuum* lines confers a new and complete resistance to pepper veinal mottle virus. Phytopathology 86:739–743. https://doi.org/10.1094/Phyto-86-739

Caranta C, Thabuis A, Palloix A (1999) Development of a CAPS marker for the Pvr4 locus: a tool for pyramiding potyvirus resistance genes in pepper. Genome 42:1111–1116. https://doi.org/10.1139/gen-42-6-1111

Chunthawodtiporn J, Hill T, Stoffel K, Van Deynze A (2018) Quantitative trait *Loci* controlling fruit size and other horticultural traits in bell pepper (*Capsicum annuum*). Plant Genome 11(1)

Coleman C, Copetti D, Cipriani G, Hoffmann S, Kozma P, Kovács L, Morgante M, Testolin R, Di Gaspero G (2009) The powdery mildew resistance gene REN1 co-segregates with an NBS-LRR gene cluster in two Central Asian grapevines. BMC Genet 10:89

D'Agostino N, Traini A, Frusciante L, Chiusano ML (2007) Gene models from ESTs (GeneModelEST): an application on the *Solanum lycopersicum* genome. BMC Bioinf 8(Suppl 1):S9–S9. https://doi.org/10.1186/1471-2105-8-S1-S9

Devran Z, Kahveci E, Özkaynak E, Studholme DJ, Tör M (2015) Development of molecular markers tightly linked to *Pvr*4 gene in pepper using next-generation sequencing. Mol Breed 35(4):101

Harris RS (2007) Improved pairwise alignment of genomic DNA. Ph.D. thesis, The Pennsylvania State University, University Park, USA

Hulse-Kemp AM, Maheshwari S, Stoffel K, Hill TA, Jaffe D, Williams S et al (2018) Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. Hort Res. https://doi.org/10.1038/s41438-017-0011-0

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucl Acids Res 45(Database issue), D353–D361. https://doi.org/10.1093/nar/gkw1092

Kang YJ, Ahn YK, Kim KT, Jun TH (2016) Resequencing of *Capsicum annuum* parental lines (YCM334 and Taean) for the genetic analysis of bacterial wilt resistance. BMC Plant Biol 16:235

Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA et al (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat Genet 46:270–278

Kim S, Park J, Yeom SI, Kim YM, Seo E, Kim KT et al (2017) New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biol 18:210. https://doi.org/10.1186/s13059-017-1341-9

Leonore R, Shabari S, Donghui L, Eva H (2017) Using the arabidopsis information resource (TAIR) to find information about arabidopsis genes. Curr Prot Bioinformat 60(1):11111–111145. https://doi.org/10.1002/cpbi36

Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J et al (2002) Apollo: a sequence annotation editor. Genome Biol 3(12). https://doi.org/10.1186/gb-2002-3-12-research0082

Li YH, Zhao SC, Ma JX, Li D, Yan L, Li J et al (2013) Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genom 14:579. https://doi.org/10.1186/1471-2164-14-579

Maumus F, Quesneville H (2016) Impact and insights from ancient repetitive elements in plant genomes. Curr Opin Plant Biol 30:41–46. https://doi.org/10.1016/jpbi201601003

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR–encoding genes in Arabidopsis. Plant Cell 15:809–834

Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. Nature 457 (7231):843–848

Qin C, Yu C, Shen Y, Fang X, Chen L, Min J et al (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proc Natl Acad Sci USA 111:5135–5140

The Gene Ontology Consortium (2017) Expansion of the gene ontology knowledgebase and resources. Nucl Acids Res 45(Database issue):D331–D338. https://doi.org/10.1093/nar/gkw1108

The UniProt Consortium (2015) UniProt: a hub for protein information. Nucl Acids Res 43(Database issue):D204–D212. https://doi.org/10.1093/nar/gku989

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 30,485(7400):635–641. https://doi.org/10.1038/nature11119

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB (2017) Direct determination of diploid genome sequences. Genome Res 27:757–767

Wu F, Eanetta NT, Xu Y, Durrett R, Mazourek M, Jahn MM, Tanksley SD (2009) A COSII genetic map of the pepper genome provides a detailed picture synteny with tomato and new insights into recent chromosome evolution in the genus *Capsicum*. Theor Appl Genet 118:1279–1293. https://doi.org/10.1007/s00122-009-0980-y

Yandell M, Ence D (2012) A beginners guide to eukaryotic genome annotation. Nat Rev Gen 13:329. https://doi.org/10.1038/nrg3174